A/B Testing

Author: Palak Sarawagi

Fixed Effect Regression

In a dataset for multiple entities across multiple time periods (also called Panel Data), a few characteristics are different for different data but are constant over time. These variables are time-invariant. Example: In a dataset for young males, their race is time-invariant.
Similarly, for the same data, there are certain characteristics that vary across time periods but not across individuals. These variables are time-invariant. Example: In a dataset for young males, the number of schools in a country is time-variant.

Why do we need a fixed-effect regression model?
A/B testing can sometimes be unethical to run, and we may have to find the causal effects using observable data. So as an alternative, if we run a regression on the variable of interest (dependent variable) and the outcome variable (independent variable), we might run into variable bias due to the unobservable time-variant and time-invariant data(described above). To remove this bias, we run the fixed effect regression model.

What does a fixed-effect regression model do?
A fixed-effect regression model
(i) removes the time-invariant independent variables replacing them with v_i which represents a unique value for each unit in the panel data (this is called the fixed effect or individual effect)
(ii) removes the time-variant independent variables replacing them with w_i

Time Dummy

The variables that change over time (time-variant) for all data can be unobservable in regressions and they can be controlled using time dummies.

This model using dummy variables allows us to eliminate bias arising from time-variant variables.

For example- Data can be observed across multiple years. Years change over time for all the data and they can be represented as time dummies in a regression to control time trends.

Instrumental Variables

Instrumental variables are used to determine causality between two variables.

Let us understand the same using an example –
To establish if Depression is caused by War participation.
A (War participation) -> B (Depression)
Is this true?

Step 1:
Introduce a third variable C (instrumental variable) that has an association with variable A but has absolutely no association with variable B (except through variable A).

Let us say our third variable C, is Draft Number which is correlated to A (War participation) but has zero association with B (Depression). The draft number is also the source of randomness or exogenous variable.

Step 2:
Use a 2-stage least square process.
Stage 1 of this process is to establish a relationship between instrument C (Draft Number) and variable A (War Participation). We run a regression for War Participation on Draft Number.

$War = \beta_0 + \beta_1 Draft + E$

Using this β_0,β_1, and data from the draft number we calculate War'.
War' -> Predicted probability of going to war based on the randomized instrument variable.

In Stage 2, we run a regression for Depression on War'. Theoretically, there should be no causal relationship between Draft Number and War Participation. But, if we see any significant β_3 in this regression, it means that this relationship is due to War Participation only. Hence, we can determine the causality.

$Depression = \beta_2 + \beta_3 War' + E$

Example:

There are 140 countries and 5 years in this dataset.

```
> #number of countries
> length(unique(homework_2$COUNTRY))
[1] 140
> #number of years
> length(unique(homework_2$YEAR))
[1] 5
```

It can be determined from the OLS model that for every 0.029 units increase in per capitaexpenditure on health, the composite measure of healthcare attainment increases by 1 unit.

```
> #model without time dummies
> ptd <-plm(LCOMP ~ LHEXP + LDALE + LHC +  LGDPC + OECD + POPDEN,
+           data = homework_2,
+           index=c("COUNTRY","YEAR"),
+           model="pooling")
> stargazer(ptd,
+           se=list(
+             sqrt(diag(vcovHC(ptd,
+                              method="arellano",
+                              type="HC1")))),
+           title="Panel OLS",
+           type="text",
+           model.numbers=FALSE,
+           column.labels=c("no time dummies"),
+           omit = c("factor[(]YEAR[)]"),
+           add.lines = list(c("Year Dummies","Yes","Yes")))

Panel OLS
==========================================
                   Dependent variable:
               ---------------------------
                          LCOMP
                     no time dummies
------------------------------------------
LHEXP                     0.029*
                          (0.015)

LDALE                    0.408***
                          (0.034)

LHC                       0.030**
                          (0.013)

LGDPC                      0.014
                          (0.017)

OECD                      0.022**
                          (0.009)

POPDEN                   -0.00000
                         (0.00000)

Constant                 2.331***
                          (0.122)

------------------------------------------
Year Dummies                Yes
Observations                700
R2                         0.913
Adjusted R2                0.912
F Statistic  1,213.604*** (df = 6; 693)
==========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

```
> #model with time dummies
> ptd <-plm(LCOMP ~ LHEXP + LDALE + LHC +  LGDPC +  OECD + POPDEN+
+             factor(YEAR),
+           data = homework_2,
+           index=c("COUNTRY","YEAR"),
+           model="pooling")
> stargazer(ptd,
+           se=list(
+             sqrt(diag(vcovHC(ptd,
+                              method="arellano",
+                              type="HC1")))),
+           title="Panel OLS",
+           type="text",
+           model.numbers=FALSE,
+           column.labels=c("time dummies"),
+           omit = c("factor[(]YEAR[)]"),
+           add.lines = list(c("Year Dummies","Yes","Yes")))

Panel OLS
==========================================
                   Dependent variable:
               ---------------------------
                          LCOMP
                     time dummies
------------------------------------------
LHEXP                     0.029*
                          (0.016)

LDALE                    0.408***
                          (0.034)

LHC                       0.030**
                          (0.013)

LGDPC                      0.013
                          (0.018)

OECD                      0.022**
                          (0.009)

POPDEN                   -0.00000
                         (0.00000)

Constant                 2.332***
                          (0.123)

------------------------------------------
Year Dummies                Yes
Observations                700
R2                         0.913
Adjusted R2                0.912
F Statistic   724.023*** (df = 10; 689)
==========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
>
```

The coefficients obtained above do not represent the causal effects because of the fixed effect variablespresent in the model as unobservable. We may observe a variable bias due to the unobservable time-variant and time-invariant data. In other words, the composite measure of healthcare attainment could be improving due to some other unobserved variable that is not considered in our model. To remove this bias, we run the fixed effect regression model.

Appropriate estimation of the causality model can be done through a first-differences model or afixed-effect model. For example, let us rerun the regression as above but in a fixed-effect model. Upon observing the result from this model (as shown below), it can be observed that per capita expenditure on health does not havesignificantly affect composite measure of healthcare attainment. Clearly, fixed model makes a lot of difference and captures the causal relationship better than OLS.

```
> #model with time dummies
> ptd <-plm(LCOMP ~ LHEXP + LDALE + LHC +  LGDPC +
+             factor(YEAR),
+           data = homework_2,
+           index=c("COUNTRY","YEAR"),
+           model="within")
> stargazer(ptd,
+           se=list(
+             sqrt(diag(vcovHC(ptd,
+                              method="arellano",
+                              type="HC1")))),
+           title="Panel OLS",
+           type="text",
+           model.numbers=FALSE,
+           column.labels=c("fixed effect"),
+           omit = c("factor[(]YEAR[)]"),
+           add.lines = list(c("Year Dummies","Yes","Yes")))

Panel OLS
=========================================
                  Dependent variable:
              ---------------------------
                         LCOMP
                     fixed effect
-----------------------------------------
LHEXP                    0.001
                        (0.001)

LDALE                   0.573***
                         (0.035)

LHC                      0.022**
                         (0.009)

-----------------------------------------
Year Dummies               Yes
Observations               700
R2                        0.930
Adjusted R2               0.911
F Statistic  1,044.673*** (df = 7; 553)
=========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```